(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
6 February 2003 (06.02.2003)

PCT

(10) International Publication Number
WO 03/010337 A1

(51) International Patent Classification[7]: C12Q 1/68, G06F 19/00, G01N 33/574

(21) International Application Number: PCT/JP01/06330

(22) International Filing Date: 23 July 2001 (23.07.2001)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicants and
(72) Inventors: **OKA, Masaaki** [JP/JP]; 3-5-8, Kitasakoshinmachi, Ube-shi, Yamaguchi 755-0093 (JP). **HAMAMOTO, Yoshihiko** [JP/JP]; 166-2, Ohaza Okiube, Ube-shi, Yamaguchi 755-0001 (JP). **OKABE, Hisafumi** [JP/JP]; 2-1-102, Katsuradaihigashi, Sakae-ku, Yokohama-shi, Kanagawa 247-0032 (JP).

(74) Agents: **SHIMIZU, Hatsushi** et al.; Kantetsu Tsukuba Bldg. 6F, 1-1-1, Oroshi-machi, Tsuchiura-shi, Ibaraki 300-0847 (JP).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 03/010337 A1

(54) **Title:** SCORING SYSTEM FOR THE PREDICTION OF CANCER RECURRENCE

(57) **Abstract:** The present invention relates to a scoring system for the prediction of cancer recurrence. More particularly, the present invention concerns with the selection of genes and/or proteins, and generation of formulae with the selected genes and/or proteins for the prediction of cancer recurrence by measuring the expression of genes and/or proteins of human tumor tissues, and comparing their patterns with those of the gene and/or protein expression of human primary tumors from patients who have cancer recurrence and those who do not have cancer recurrence. The present invention also relates to a kit for performing the method of the present invention comprising DNA chip, oligonucleotide chip, protein chip, peptides, antibodies, probes and primers that are necessary for effecting DNA microarrays, oligonucleotide microarrays, protein arrays, northern blotting, in situ hybridization, RNase protection assays, western blotting, ELISA assays, reverse transcription polymerase-chain reaction (hereinafter referred to as RT-PCR) to examine the expression of at least 2 or more of genes and/or proteins, preferably 4 or more of genes and/or proteins, more preferably 6 or more of genes and/or proteins, and most preferably 12 or more of genes and/or proteins, that are indicative of cancer recurrence.

B1

Description

Scoring System for the Prediction of Cancer Recurrence

The present invention relates to a scoring system for the prediction of cancer recurrence. More particularly, the present invention concerns with the selection of genes and/or proteins, and generation of formulae with the selected genes and/or proteins for the prediction of cancer recurrence by measuring the expression of genes and/or proteins of human tumor tissues, and comparing their patterns with those of the gene and/or protein expression of human primary tumors from patients who have cancer recurrence and those who do not have cancer recurrence.

The present invention also relates to a kit for performing the method of the present invention comprising DNA chip, oligonucleotide chip, protein chip, peptides, antibodies, probes and primers that are necessary for effecting DNA microarrays, oligonucleotide microarrays, protein arrays, northern blotting, in situ hybridization, RNase protection assays, western blotting, ELISA assays, reverse transcription polymerase-chain reaction (hereinafter referred to as RT-PCR) to examine the expression of at least 2 or more of genes and/or proteins, preferably 4 or more of genes and/or proteins, more preferably 6 or more of genes and/or proteins, and most preferably 12 or more of genes and/or proteins, that are indicative of cancer recurrence.

Background of the invention

Cancer is one of the major causatives of death in the world. The overall prevalence rate of cancer is about 1 % of the population and yearly incidence rate is about 0.5 %. About one out of ten patients discharged from hospitals have cancer as their primary diagnosis. The main existing treatment modalities are surgical resection, radiotherapy, chemotherapy, and biological therapy including hormonal therapy. Furthermore, newly developed biotechnologies have been offering new treatment modalities, such as gene therapy. Nevertheless, cancer is dreaded disease because in most cases there is no really effective treatment available. One of the major difficulties of cancer treatment is the ability of cancer cells to become resistant to drugs and to spread to other sites of tissues, where they can generate new tumors, which often results in recurrence. If a cancer recurrence is predictable before recurrence occurs, such cancer becomes curable by local treatment with surgery.

Among various tumors, hepatocellular carcinoma (hereinafter referred to as HCC) is one of the most common fatal cancers in the world and the number of incidences is increasing in many countries including the USA, Japan, China and European countries. Both hepatitis B virus (hereinafter referred to as HBV) and hepatitis C virus (hereinafter referred to as HCV) infections can be a causative of HCC. In fact, increase in HCC patients is in parallel to an increase in chronic HCV infection (El-Serag, H.B. & Mason, A.C. Rising incidence of hepatocellular carcinoma in the United States, *N. Engl. J. Med.* **340,** 745-750 (1999) and Okuda, K. Hepatocellular carcinoma, *J. Hepatol.* **32,** 225-237 (2000)). Despite the elevated incidences of HCC, there is no promising therapy for this disease. The major problem in the treatment of HCC is intrahepatic metastasis. Recurrence was observed in 30 to 50% of HCC patients who had received hepatic resection (Iizuka, N. *et al.* NM23-H1 and NM23-H2 messenger RNA abundance in human hepatocellular carcinoma, *Cancer Res.* **55,** 652-657 (1995), Yamamoto, J. *et al.* Recurrence of hepatocellular carcinoma after surgery, *Br. J. Surg.* **83,** 1219-1222 (1996), and Poon, R.T. *et al.* Different risk factors and prognosis for early and late intrahepatic recurrence after resection of hepatocellular carcinoma, *Cancer* **89,** 500-507 (2000)). Although the pathologic TNM staging system has been applied in the treatment of HCC, this system is poorly predictive of recurrences in patients who undergo hepatic resection (Izumi, R. *et al.* Prognostic factors of hepatocellular carcinoma in patient undergoing hepatic resection, *Gastroenterology* **106,** 720-727 (1994)). A number of molecules have also been proposed as predictive markers for HCCs, none of them has proven to be clinically useful (Iizuka, N. *et al.* NM23-H1 and NM23-H2 messenger RNA abundance in human hepatocellular carcinoma, *Cancer Res.* **55,** 652-657 (1995), Hsu, H.C. *et al.* Expression of p53 gene in 184 unifocal hepatocellular carcinomas: association with tumor growth and invasiveness, *Cancer Res.* **53,** 4691-4694 (1993), and Mathew, J. *et al.* CD44 is expressed in hepatocellular carcinomas showing vascular invasion, *J. Pathol.* **179,** 74-79 (1996)). Thus, any method to predict recurrence would be quite valuable to understand cancer mechanisms and also to establish the new therapies for cancer. However, because there are technological limitations for predicting recurrence by the traditional methods and further limitations may be attributable to high inter-patient heterogeneity of tumors, it is necessary to devise a novel method to characterize tumors and predict cancer recurrence.

Recent development of microarray technologies, which allow one to perform parallel expression analysis of a large number of genes, has opened up a new era in medical science (Schena, M. *et al.* Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science* **270,** 467-470 (1995), and DeRisi, J. *et al.* Use of a cDNA microarray to analyze gene expression patterns in human cancer, *Nature Genet.* **14,** 457-460 (1996)). In particular, studies by cDNA microarrays of the gene expression of tumors have provided significant

insights into the properties of malignant tumors such as prognosis and drug-sensitivity (Alizadeh, A.A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature* **403**, 503-511 (2000), and Scherf, U. *et al.* A gene expression database for the molecular pharmacology of cancer, *Nature Genet.* **24**, 236-244 (2000)).

Recently, supervised learning has been introduced into gene-expression analysis (Brazma, A. & Vilo, J. Gene expression data analysis, *FEBS Lett.* **480**, 17-24 (2000) and Kell, D.B. & King, R.D. On the optimization of classes for the assignment of unidentified reading frames in functional genomics programs: the need for machine learning, *Trends Biotechnol.* **18**, 93-98 (2000)). Using classified samples, supervised learning has the conclusive advantage of much a priori knowledge about the nature of the data (Duda, R.O. *et al. Pattern classification*, John Wiley & Sons (2001), and Jain, A.K. *et al.* Statistical pattern recognition: A review, *IEEE Trans. Pattern Analysis and Machine Intelligence.* **22**, 4-37 (2000)). However, none of supervised learning methods previously published directly evaluates the combination of genes and thus can utilize information concerning the statistical characteristics, i.e., structure of the distribution of genes (Golub, T.R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* **286**, 531-537 (1999), and Brown, M.P. *et al.* Knowledge-based analysis of microarray gene expression data by using support vector machines, *Proc. Natl. Acad. Sci. USA* **97**, 262-267 (2000)).

Scoring systems that are predictive of cancer recurrence are created by analyzing the DNA microarray data with supervised learning in statistical pattern recognition (Duda, R.O. *et al. Pattern classification,* John Wiley & Sons (2001)).

Supervised learning in statistical pattern recognition has been successfully applied to resolve a variety of issues such as document classification, speech recognition, biometric recognition, and remote sensing (Jain, A.K. *et al.* Statistical pattern recognition: A review, *IEEE Trans, Pattern Analysis and Machine Intelligence.* **22**, 4-37 (2000)).

In the present invention, the inventors provide a scoring system to predict cancer recurrence by analyzing the expression of genes and/or proteins of human primary tumors. That is the invention concerns a method for the prediction of cancer recurrence which comprises measuring the expression of genes and/or proteins of human tumor tissues, and comparing it with the expression of the genes and/or proteins of human primary tumors from patients who have cancer

recurrence and those who do not have cancer recurrence.

Brief Description of the Drawings

Figure 1 illustrates the procedure of gene selection (Steps 1-7) and evaluation (Steps 8-10) of the scoring system with the optimal gene subset.

Figure 2 illustrates the optimal number of genes.

Figure 3 illustrates the average differences of the mRNA for the genes selected for the prediction of early intrahepatic recurrence. The average differences of the mRNA for the 12 genes were compared between Group A (indicated as A) and Group B (indicated as B).

Figure 4 illustrates the relation between virus type, TNM stage, and scores (T values) for the prediction of early intrahepatic recurrence. Using the optimal subset of 12 genes, the scoring system created with 30 training samples was evaluated with 3 test samples. This operation was independently repeated 10 times. The T values for all of the test sample were caluculated. Early intrahepatic recurrence was predicted when the T value is below zero. Regardless of stage and virus types, all HCCs with a negative T value had early intrahepatic recurrences and all HCCs with a positive T value had no recurrences. Filled, Group A (patients with early intrahepatic recurrence); White, Group B (patients without early intrahepatic recurrence); ○, stage I;◇, stage II; △, stage IIIA; □, stage IVA. B; HBV-positive, C; HCV-positive, N; HBV- HCV-double negative.

Figure 5 illustrates the scoring system.

Detailed explanation of the invention

In the present invention, human tissues from tumors including those of brain, lung, breast, stomach, liver, pancreas, gallbladder, colon, rectum, kidney, bladder, ovary, uterus, prostate, and skin are used. After human tissues are resected during surgeries, it is preferable that they are immediately frozen in liquid nitrogen or acetone containing dry ice and stored at between -70 and -80°C until use with or without being embedded in O.C.T. compound (Sakura-Seiki, Tokyo, Japan, Catalog No. 4583).

Expression of genes and/or proteins of tumor tissues from patients who are tested for the probability of cancer recurrence are analyzed by measuring the levels of RNA and/or proteins. In many cases,

the levels of RNA and/or proteins are determined by measuring fluorescence from substances including fluorescein and rhodamine, chemiluminescence from luminole, radioactivities of radioactive materials including ^{3}H, ^{14}C, ^{35}S, ^{33}P, ^{32}P, and ^{125}I, and optical densities. Expression levels of RNA and/or proteins are determined by known methods including DNA microarray (Schena, M. *et al.* Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science **270**, 467-470 (1995), and Lipshutz, R.J. *et al.* High density synthetic oligonucleotide arrays, *Nature Genet.* **21**, 20-24 (1999)), RT-PCR (Weis, J.H. *et al.* Detection of rare mRNAs via quantitative RT-PCR, *Trends Genetics* **8**, 263-264 (1992), and Bustin, S.A. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays, *J. Mol. Endocrinol.* **25**, 169-193 (2000)), northern blotting and in situ hybridization (Parker, R.M. & Barnes, N.M. mRNA: detection in situ and northern hybridization, *Methods Mol. Biol.* **106**, 247-283 (1999)), RNase protection assay (Hod, Y.A. Simplified ribonuclease protection assay, *Biotechniques* **13**, 852-854 (1992), Saccomanno, C.F. *et al.* A faster ribonuclease protection assay, *Biotechniques* **13**, 846-850 (1992)), western blotting (Towbin, H. *et al.* Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets, *Proc. Natl. Acad. Sci. U S A* **76**, 4350-4354 (1979), Burnette, W.N. Western blotting: Electrophoretic transfer of proteins form sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radioiodinated protein A, *Anal. Biochem.* **112**, 195-203 (1981)), ELISA assays (Engvall, E. & Perlman, P. Enzyme-linked immunosorbent assay (ELISA): Quantitative assay of immunoglobulin G, *Immunochemistry* **8**: 871-879 (1971)), and protein arrays (Merchant, M. & Weinberger, S.R. Review: Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry, *Electrophoresis* **21**, 1164-1177 (2000), Paweletz, C.P. *et al.* Rapid protein display profiling of cancer progression directly from human tissue using a protein biochip, *Drug Development Research* **49**, 34-42 (2000)).

Expression of genes and/or proteins of tumors from cancer patients who have early recurrence and those who do not are determined in the same way as that for the patients who are tested for the probability of recurrence.

Although early recurrence of cancer varies among different cancer types, it usually occurs within one or two years after resection. Therefore, tumors from cancer patients who have recurrence within one or two years after resection can be used as the tumors of patients with early recurrence, and those from patients who do not have recurrence before one or two years after

resection can be used as the tumors of patients without early recurrence.

Differences in the expression levels or patterns of genes and/or proteins of tumors between cancer patients who have early recurrence and who do not can be analyzed and detected by known methods of statistical analyses. Supervised learning in statistical pattern recognition can be used for statistical analysis of the expression patterns of genes and/or proteins of tumors. By supervised learning in statistical pattern recognition, 2 or more of genes and/or proteins of which expression is indicative of cancer recurrence are selected from the examined genes and/or proteins.

Some genes and/or proteins that are indicative of cancer recurrence are first selected by one-dimenstional criteria. Then, the optimal subsets of genes and/or proteins are selected out of these genes and/or proteins by an exhaustive search with the leave-one-out method that can take all the possible combinations of genes and/or proteins into account.

Formulae that are predictive of cancer recurrence are created by using the optimal subsets of at least 2 or more of genes and/or proteins, preferably 4 or more of genes and/or proteins, more preferably 6 or more of genes and/or proteins, and most preferably 12 or more of genes and/or proteins of which expression is indicative of cancer recurrence. Simple classifiers such as linear classifier (Duda, R.O. *et al. Pattern classification*, John Wiley & Sons (2001), and Jain, A.K. *et al.* Statistical pattern recognition: A review, *IEEE Trans. Pattern Analysis and Machine Intelligence.* **22,** 4-37 (2000)) that work well even if the number of samples is small compared to the number of genes and/or proteins are used to create formulae.

The present invention also concerns kits to carry out the methods of the present invention. Kits to examine the expression patterns of 2 or more of genes and/or proteins that are indicative of cancer recurrence consist of the components including reagents for RNA extraction, enzymes for the syntheses of cDNA and cRNA, DNA chip, oligonucleotide chip, protein chip, probes and primers for the analyses, DNA fragments of control genes, and antibodies to various proteins. Components of the kits are easily available from the market. For instance, oligonucleotide chips, guanidine-phenol reagent, reverse transcriptase, T7 RNA polymerase and taq polymerase can be purchased and assembled for the kits of the present invention.

The following examples merely illustrate the preferred method for the prediction of cancer recurrent of the present invention and are not to be construed as being limited thereto.

Examples

Example 1. Selection of the patients for analysis of early intrahepatic recurrence

It has been reported that early intrahepatic recurrences (within one year) after surgery arise mainly from intrahepatic metastases, whereas late recurrences are more likely to be multicentric occurrence (Poon, R.T. *et al.* Different risk factors and prognosis for early and late intrahepatic recurrence after resection of hepatocellular carcinoma, *Cancer* **89,** 500-507 (2000)). Moreover, it is well known that the outcome of patients with intrahepatic recurrence was worse than that of patients with multicentric occurrence (Yamamoto, J. *et al.* Recurrence of hepatocellular carcinoma after surgery, *Br. J. Surg.* **83,** 1219-1222 (1996), and Poon, R.T. *et al.* Different risk factors and prognosis for early and late intrahepatic recurrence after resection of hepatocellular carcinoma, *Cancer* **89,** 500-507 (2000)). Therefore gene-expression patterns linked to early intrahepatic recurrence were investigated within one year after surgery.

Thirty-three patients underwent surgical treatment for HCC in Yamaguchi University Hospital between May 1997 and January 2000. Informed consent in writing was obtained from all cases before surgery. The study protocol was approved by the Institutional Review Board for Human Use at the Yamaguchi University School of Medicine in May 1996. A histopathological diagnosis of HCC was made in all patients after surgery. The histopathological examination also revealed no residual tumors (R0) in all of the 33 HCC samples. Table 1 shows the clinicopathologic characteristics of the 33 patients, based on the TNM classification of Union Internationale Contre le Cancer (UICC) (Sobin, L.H. & Wittekind, C. TNM classification of Malignant Tumors, *5th ed.*, UICC, Wiley-Liss, 74-77 (1997)). Serologically, 7 patients were hepatitis B surface antigen-positive, 22 patients were anti-HCV antibody-positive, and the remaining 4 patients were negative for both. The 33 patients were tracked for cancer recurrence with ultrasonography, computed tomography, and alpha-fetoprotein level every 3 months following hepatic resection. Whenever necessary, magnetic resonance imaging and hepatic angiography were added. Of the 33 HCC patients, early intrahepatic recurrences were found in 12 (36%). In 11 of the 12 patients, recurrent HCCs were detected as multiple nodules or diffuse dissemination in the remnant liver. In one patient, a novel tumour was detected as single nodule in the segment adjacent to the resected primary lesion 9 month after surgery, and then multiple lung metastases were observed. None of the remaining 21 patients had intrahepatic recurrences and other distant metastases within one year after surgery. These patients were divided into two groups; the patients who had intrahepatic recurrences within one year in Group A (n=12) and those who did not in Group B (n=21) (Table 1). The χ^2 test and Fisher's exact test were used to elucidate differences in clinicopathologic factors between the 2 groups.

Example 2. Extraction of the RNA from tissues

Pieces of the tissues (about 125mm^3) were suspended in TRIZOL (Life Technologies, Gaithersburg, USA, Catalog No. 15596-018) or Sepasol-RNAI (Nacalai tesque, Kyoto, Japan, Catalog No. 306-55) and homogenized twice with a Polytron(Kinematica, Littau, Switzerland) (5 sec. at maximum speed). After addition of chloroform, the tissues homogenates were centrifuged at 15,000 x g for 10 min, and aqueous phases, which contained RNA, were collected. Total cellular RNA was precipitated with isopropyl alcohol, washed once with 70% ethanol and suspended in DEPC-treated water (Life Technologies, Gaithersburg, USA, Catalog No. 10813-012). After RNA was treated with 1.5 units of DNase I (Life Technologies, Gaithersburg, USA, Catalog No. 18068-015), the RNA was re-extracted with TRIZOL/chloroform, precipitated with ethanol and dissolved in DEPC-treated water. Thereafter, small molecular weight nucleotides were removed by using RNeasy Mini Kit (QIAGEN, Hilden, Germany, Catalog No. 74104) according to a manufacture's instruction manual. Quality of the total RNA was judged from ratio of 28S and 18S ribosomal RNA after agarose gel electrophoresis. The purified total RNA was stored at –80 °C in 70% ethanol solution until use.

Example 3. Synthesis of cDNA and labeled cRNA probes

cDNA was synthesized by using reverse SuperScript Choice System (Life Technologies, Gaithersburg, USA, Catalog No. 18090-019) according to the manufacture's instruction manual. Five microgram of the purified total RNA was hybridized with an oligo-dT primer (Sawady Technology, Tokyo, Japan) that contained the sequences for the T7 promoter and 200 units of SuperScriptII reverse transcriptase and incubated at 42 °C for 1 hr. The resulting cDNA was extracted with phenol/chloroform and purified with Phase Lock Gel Light (Eppendorf, Hamburg, Germany, Catalog No. 0032 005.101).

cRNA was also synthesized by using MEGAscript T7 kit (Ambion, Austin, USA, Catalog No. 1334) and the cDNA as templates according to the manufacture's instruction. Approximately 5 μg of the cDNA was incubated with 2 μl of enzyme mix containing T7 polymerase, 7.5 mM each of adenosine triphosphate (ATP) and guanosine triphosphate (GTP), 5.625 mM each of cytidine triphosphate (CTP) and uridine triphosphate (UTP), 1.875 mM each of Bio-11-CTP and

Bio-16-UTP (ENZO Diagnostics, Farmingdale, USA, Catalog No. 42818 and 42814, respectively) at 37 °C for 6 hr. Mononucleotides and short oligonucleotides were removed by column chromatography on CHROMA SPIN +STE-100 column (Clontech, Palo Alto, USA, Catalog No. K1302-2), and the cRNA in the eluates was sedimented by adding ethanol. Quality of the cRNA was judged from the length of the cRNA after agarose gel electrophoresis. The purified cRNA was stored at –80 °C in 70% ethanol solution until use.

Example 4. Gene expression analysis of tumors from patients with and without recurrence

Gene expression of human primary tumors from live cancer patients were examined by high-density oligonucleotide microarrays (HuGeneFL array, Affymetrix, Santa Clara, USA, Catalog No. 510137) (Lipshutz, R.L. *et al.* High density synthetic oligonucleotide arrays, *Nature Genet.* **21,** 20-24 (1999)). For hybridization with oligonucleotides on the chips, the cRNA was fragmented at 95 °C for 35 min in a buffer containing 40 mM Tris (Sigma, St. Louis, USA, Catalog No. T1503)-acetic acid (Wako, Osaka, Japan, Catalog No. 017-00256) (pH8.1), 100 mM potassium acetate (Wako, Osaka, Japan, Catalog No. 160-03175), and 30mM magnesium acetate (Wako, Osaka, Japan, Catalog No. 130-00095). Hybridization was performed in 200μl of a buffer containing 0.1M 2-(N-Morpholino) ethanesulfonic acid (MES) (Sigma, St. Louis, USA, Catalog No. M-3885) (pH6.7), 1M NaCl (Nacalai tescque, Tokyo, Japan, Catalog No. 313-20), 0.01% polyoxylene(10) octylphenyl ether (Wako, Osaka, Japan, Catalog No. 168-11805), 20 μg herring sperm DNA (Promega, Madison, USA, Catalog No. D181B), 100μg acetylated bovine serum albumin (Sigma, St. Louis, USA, Catalog No. B-8894), 10 μg of the fragmented cRNA, and biotinylated-control oligonucleotides, biotin-5'-CTGAACGGTAGCATCTTGAC-3' (Sawady technology, Tokyo, Japan) at 45 °C for 12 hr. After washing the chips with a buffer containing 0.01M MES (pH6.7), 0.1M NaCl, 0.001% polyoxylene(10) octylphenyl ether buffer, the chips were incubated with biotinylated anti-streptavidin antibody (Funakoshi, Tokyo, Japan, Catalog No. BA0500) and staining with streptavidin R-Phycoerythrin (Molecular Probes, Eugene, USA, Catalog No. S-866) to increase hybridization signals as described in the instruction manual (Affymetrix, Santa Clara, USA). Each pixel level was collected with laser scanner (Affymetrix, Santa Clara, USA) and levels of the expression of each cDNA and reliability (Present/Absent call) were calculated with Affymetrix GeneChip ver.3.3 and Affymetrix Microarray Suite ver.4.0 softwares. From this experiments, expression of 6000 genes in the human primary tumors of liver cancer patients are determined.

Example 5. Kinetic RT-PCR analysis

Expression of genes is also determined by kinetic RT-PCR. Kinetic RT-PCR was performed by a real-time fluorescence PCR system. PCR amplification using a LightCycler instrument (LightCycler system, Roche Diagnostics, Mannheim, Germany, Catalog No. 2011468) was carried out in 20 μl of reaction mixture consisting of a master mixture and buffer (LightCycler DNA Master hybridization probes, Roche Diagnostics, Mannheim, Germany, Catalog No. 2158825), 4 mM magnesium chloride (Nacalai tescque, Tokyo, Japan, Catalog No. 7791-18-6), 10 pmoles of PCR primers (Sawady Technology, Tokyo, Japan), 4 pmoles of fluorescent hybridization probes (Nihon Genome Research Laboratories, Sendai, Japan), which were designed to hybridize with the target sequences in a head-to-tail arrangement on the strand of amplified products, and 2 μl of template cDNA in a LightCycler capillary (Roche Diagnostics, Mannheim, Germany, Catalog No. 1909339). The donor probes was labeled at the 3'-end with fluorescence, while the acceptor probe was labeled at the 5'-end with LC-Red640 and modified at the 3'- end by phosphorylation to block extension. The gap between the 3'-end of the donor probe and the 5'-end of the acceptor probe was between 1 and 3 bases. Prior to amplification, 0.16 μl of TaqStart antibody (Clontech, Palo Alto, USA, Catalog No. 5400-1) was added to the reaction mixture, which was followed by the incubation at room temperature for 10 min to block primer elongation. Then, the antibody was inactivated by the incubation at 95°C for 90 sec., and the amplification was performed in the LightCycler by 40 cycles of incubation at 95 °C for 0 sec. for denaturation, at 57-60 °C for 3-10 sec. for annealing and at 72 °C for 10 sec. for extension, with a temperature slope of 20 °C/sec. Real-time PCR monitoring was achieved by measuring the fluorescent signals at the end of the annealing phase in each amplification cycle. To qualify the integrity of isolated RNA and normalize the copy number of target sequences, kinetic RT-PCR analysis for glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was also carried out by using hybridization probes. External standards for the target mRNA and GAPDH mRNA were prepared by 10-fold serial dilutions (10^3 to 10^8) of plasmid DNA. Quantification of mRNA in each sample was performed automatically by reference to the standard curve constructed at each time point according to the LightCycler software (LightCycler software version 3, Roche Diagnostics, Mannheim, Germany).

Example 6. Identification of sets of genes of which expression distinguishes the liver cancer patients who have early intrahepatic recurrence from those the patients who do not have early intrahepatic recurrence

Early intrahepatic recurrence tended to be associated with the number of primary tumor and TNM stage with the p values of 0.041 and 0.006, respectively, but not with the other clinicopathologic factors (Table 1). The number of primary tumors at the time of surgery distinguished group A from group B only with the limited sensitivity and specificity (62 % and 75 %, respectively). The TNM staging also had a limited sensitivity (67 %) and specificity (83 %) for the separation of groups A and B. Thus, it appears that these traditional classifications cannot be predictive of the early intrahepatic recurrence.

Supervised learning in statistical pattern recognition was applied to analyze the data of high-density oligonucleotide microarrays. The scoring system was designed with the training samples and was validated its performance with the test samples (Fig. 1). In order to maintain independence of the training and test samples, the cross-validation approach in which the training and the test samples were interchanged was adopted. Thirty-three available samples were devided into 30 training samples and 3 test samples by the cross-validation approach (Fig. 1, Step 1). On the basis of a prior probability, ten sets of the training samples consisting of 11 samples from Group A and 19 samples from Group B were created. As a result, ten sets of three test samples consisting of one from Group A and two from Group B were created.

Fifty useful genes were selected to create the predictive scoring system from all the examined genes that had mean average differences of more than twofold between Group A and B using the Fisher criterion (Fig. 1, Steps 2-3), which was given by the following Formula (I),

$$F(i) = \frac{(\mu_A(i) - \mu_B(i))^2}{P(A)\sigma_A^2(i) + P(B)\sigma_B^2(i)}$$

where $\mu_A(i)$ is the i th component of the sample mean vector μ_A of Group A, $\sigma_A^2(i)$ is the i th diagonal element of the sample covariance matrix Σ_A of Group A, and $P(A)$ is the a priori probability of Group A.

Then, the optimal subset of the genes for the scoring system was identified as mentioned below.

The Fisher linear classifier assigns a test sample x to be classified to Group A in the following Formula (II).

if $F_A(x) < F_B(x)$

where

$$F_A(x) = \frac{1}{2}(x-\mu_A)^T \Sigma_W^{-1}(x-\mu_A) - \ln P(A)$$
$$\Sigma_w = P(A)\Sigma_A + P(B)\Sigma_B$$

In the leave-one-out method, the sample mean vector, sample covariance matrix, and the a priori probability were estimated by using 29 samples as training samples. Then, the resulting Fisher linear classifier was testd on the remaining sample as a pseudo-test sample. This operation was repeated 30 times. The error rate was calculated for each possible subset of the genes. For example, when selecting 5 genes out of 50, the number of subsets to be examined is two million.

Next, candidate gene subsets minimizing the error rate were selected (Fig. 1, Step 4). This trial was independently repeated 10 times (Fig. 1, Step 5).

Among the candidate gene subsets, the gene subset that most frequently appeared throughout the 10 trials was selected as the optimal subset of the genes for the discrimination of the two groups (Fig. 1, Step 6). Using the optimal subset of genes selected, the score T is given by the following Formula (III).

$$T(x) = F_A(x) - F_B(x)$$

In this scoring system, all HCCs with a negative T value are classified into Group A (early intrahepatic recurrence group) and all HCCs with a positve T value are classified into Group B (nonrecurrence group).

The optimal number of the genes was determined according to the criterion J that was given by the following formula (IV) (Fig. 1, Step 7).

$$J = \frac{1}{30}\left[\sum_{x \in B} T(x) - \sum_{x \in A} T(x)\right]$$

The criterion J measures the separability of Group A from B. The average and 95% confidence interval of the J values in 10 different training sets were computed for various numbers of the genes (Fig. 2). The separability became better in parallel to an increase in the number of the genes. Ninety-five percentage of the confidence interval became almost similar when the number of the

genes reached 10 and 12, indicating that the 12 is the most appropriate number of the genes for the separability of the two groups (Fig. 2).

Example 7. The optimal subset of the 12 genes of which expression is indicative of early intrahepatic recurrence

According to the algorithm described above, the optimal subset of the 12 genes that discriminates Group A from Group B was identified. The optimal gene subset consisted of the genes for platelet-derived growth factor receptor alpha (PDGFRA), tumor necrosis factor alpha (TNF-α) inducible protein A20, lysosomal-associated multitransmembrane protein (LAPTm5), HLA-DR alpha heavy chain, rel proto-oncogene, Staf50, putative serine/threonine protein kinase, MADS/MEF2-family transcription factor (MEF2C), HUMLUCA19 Human cosmid clone LUCA19 from 3p21.3, DEAD-box protein p72, vimentin and KIAK0002 (Table 2). Of the 12 genes selected, expression of the eleven were down-regulated in Group A; the mean of the average differences of these genes in Group A were less than half of those in Group B (Fig. 3). In contrast, the HUMLUCA19 gene expression was up-regulated in Group A; the mean of the average differences of the HUMLUCA19 gene in Group A was increased by more than 3-fold compared to that in Group B (Fig. 3). Accuracy of the scoring for the prediction of the early intrahepatic was evaluated with the 10 different sets of 3 test samples (Fig. 4). Early recurrence of HCC is predicted by calculating the T values of the 12 genes from HCC patients. Recurrence within one year after surgery is very likely when the T value is below zero, and recurrence within one year after surgery is quite unlikely when the T value is above zero. The scoring system could perfectly predict early intrahepatic recurrence of 3 test samples in all 10 trials (Fig. 4). The scoring system was independent of viral infection patterns and was much more accurate than TNM staging system (Fig. 4). Scoring system based on all 33 HCCs with the above 12 genes (Fig. 5) includes the following formula (V).

Formula (V)

$$T(x) = 0.053862\,x_1 + 0.038848x_2 + 0.030176x_3 + 0.001824x_4 + 0.096997x_5 + 0.017259x_6 + 0.015908x_7 + 0.103081x_8 - 0.093746x_9 + 0.024031x_{10} - 0.005417x_{11} - 0.119177x_{12} - 11.046007,$$

where x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12} are the normalized average differences of the mRNAs for platelet-derived growth factor receptor alpha (PDGFRA), tumor necrosis factor alpha (TNF-α) inducible protein A20, lysosomal-associated multitransmembrane protein (LAPTm5), HLA-DR alpha heavy chain, rel proto-oncogene, Staf50, putative serine/threonine protein kinase,

MADS/MEF2-family transcription factor (MEF2C), HUMLUCA19 Human cosmid clone LUCA19 from 3p21.3, DEAD-box protein p72, vimentin and the KIAK0002 gene (Table 2).

The 12 genes selected by the present invention are involved in a wide range of biological processes. Of these, immune response-related genes such as HLA-DR alpha heavy chain, TNF-α inducible protein A20 and Staf50, were down-regulated in HCCs with early intrahepatic recurrence. Because HLA-DR alpha heavy chain is considered to play an important role in the antigen-presenting by macrophages (Tissot, C. & Mechti, N. Molecular cloning of a new interferon-induced factor that represses human immunodeficiency virus type 1 long terminal repeat expression, *J. Biol. Chem.* **270,** 14891-14898 (1995)), its down-regulation in tumorous tissues might facilitate escape of tumor cells from host immune surveillance. Rel proto-oncogene, which is involved in intracellular signaling pathway as well as NF-κB, was also down-regulated in HCCs with early intrahepatic recurrence. Furthermore, the expression of rel/NF-κB have been reported to be associated with T-cell activation (Mora, A. *et al.* NF-kappa B/Rel participation in the lymphokine-dependent proliferation of T lymphoid cells, *J. Immunol.* **166,** 2218-2227 (2000)). Thus, it seems that several genes that were selected for the use to predict early intrahepatic recurrence by the present invention are involved in the weakening the host immune responses against HCC cells possessing high metastatic potentials.

Gene expression pattern of other HCC patients whose follow-up period recently reached one year was also analyzed by oligonucleotide microarray, and the scores of the expression of 12 genes were calculated according to the formula described above. T values of patients who lived without recurrence more than one year after surgery were positive (plus score) and that of the other patient who had intrahepatic recurrence within one year after surgery was negative (minus). Thus, the scoring system consisting of the subset of 12 genes obtained from 6000 could predict early intrahepatic recurrence accurately. The application of supervised learning in statistical pattern recognition to clinical specimens may provide a key information in advances for prevention, diagnosis, and therapeutics of other diseases as well as HCC. Furthermore, not only DNA microarray but also other methods such as RT-PCR can be used to determine the expression of the optimal sets of genes.

Table 1

Clinicopathologic factors of the HCCs used to the early intrahepatic recurrence.

Factors	Group A (n =12)	Group B (n =21)	P value
Sex			N.S.
Male	8	16	
Female	4	5	
Age			N.S.
≦60	5	7	
>60	7	14	
Viral infection			N.S.
HBV	3	4	
HCV	8	14	
Non B,Non C	1	3	
Primary lesion			0.041
Single tumor	3	13	
Multipe tumors	9	8	
Tumor size (cm)			N.S.
<2.0	0	5	
2.0-5.0	8	14	
>5.0	4	2	
Stage*			0.006
I/II	2	14	
IIIA/IVA	10	7	
Histological grading*			N.S.
G1	0	2	
G2	9	17	
G3	3	2	
Venous invasion*			N.S.
(-)	7	18	
(+)	5	3	
Non-tumorous liver			N.S.
Non-specific change	1	1	
Chronic hepatitis	2	10	
Liver cirrhosis	9	10	

*, Assessment based on TNM classification of UICC

HBV: hepatitis B virus, HCV: hepatitis C virus, non-B non-C: neither HBV nor HCV

Group A: early intrahepatic recurrence (+), Group B: early intrahepatic recurrence (-)

N.S.: Not significant

Table 2

The formula and the 12 genes to predict early intrahepatic recurrence.

Formula

$$T(x) = 0.053862x_1 + 0.038848x_2 + 0.030176x_3 + 0.001824x_4 + 0.096997x_5 + 0.017259x_6 + 0.015908x_7 + 0.103081x_8 - 0.093746x_9 + 0.024031x_{10} - 0.005417x_{11} - 0.119177x_{12} - 11.046007$$

	GB*	Description
x_1;	M21574	platelet-derived growth factor receptor alpha (PDGFRA)
x_2;	M59465	tumor necrosis factor alpha inducible protein A20
x_3;	U51240	lysosomal-associated multitransmembrane protein (LAPTm5)
x_4;	X00274	HLA-DR alpha heavy chain (class II antigen)
x_5;	X75042	rel proto-oncogene
x_6;	X82200	Staf50
x_7;	Y10032	putative serine/threonine protein kinase
x_8;	L08895	MADS/MEF2-family transcription factor (MEF2C)
x_9;	AC000063	HUMLUCA19 Human cosmid clone LUCA19 from 3p21.3
x_{10};	U59321	DEAD-box protein p72
x_{11};	Z19554	vimentin
x_{12};	D13639	KIAK0002 gene

GB*: Gene bank access number

Claims

1. A scoring system for the prediction of cancer recurrence using 2 or more genes and/or proteins selected by statistical analyses based on expression levels or patterns of genes and/or proteins of cancer tissues from human cancer patients who have recurrence and those who do not have recurrence.
2. The scoring system according to claim 1, wherein the number of genes and/or proteins selected by statistical analyses is 4 or more.
3. The scoring system according to claim 1, wherein the number of genes and/or proteins selected by statistical analyses is 6 or more.
4. The scoring system according to claim 1, wherein the number of genes and/or proteins selected by statistical analyses is 12 or more.
5. The scoring system according to claims 1-4, wherein the cancer tissues are liver cancer tissues.
6. The scoring system according to claims 1-5, wherein the expression of genes and/or proteins of human cancer tissues and the expression of genes and/or proteins of human primary cancer tissues from patients who have recurrence and those who do not have recurrence are examined by means of DNA microarray, reverse transcription polymerase-chain reaction or protein array.
7. A kit for carrying out the scoring system according to claims 1-6 comprising DNA chip, oligonucleotide chip, protein chip, probes or primers that are necessary for effecting DNA microarrays, oligonucleotide microarrays, protein arrays, northern blotting, RNase protection assays, western blotting, and reverse transcription polymerase-chain reaction to examine the expression of genes and/or proteins selected by the scoring system according to claims 1-6.
8. A method for the prediction of cancer recurrence, comprising the steps of:
   (a) examining the expression levels or patterns of genes and/or proteins in the samples prepared from the cancer tissues of patients, wherein said genes and/or proteins are selected by the scoring system according to claims 1-6; and,
   (b) predicting cancer recurrence of patients by applying the expression levels or patterns of genes and/or proteins examined in step (a) to the scoring system according to claims 1-6.

Fig. 1

Step 1. Divide 33 available samples into 30 training samples and 3 test samples by the cross-validation method.

Step 2. Compute the Fisher criterion using 30 training samples.

Step 3. Select superior 50 genes for discrimination between Groups A and B based on both the Fisher criterion and fold-change.

Step 4. Select the candidate gene subsets out of 50 genes by the exhaustive search with the leave-one-out method.

Step 5. Repeat Steps 1-4 independently (10 times).

Step 6. Select the optimal gene subset which most frequently appears in 10 trials.

Step 7. Determine the optimal number of the genes according to the criterion J computed with the 10 different sets of the training samples.

Step 8. Design the scoring system using 30 training samples selected in Step 1.

Step 9. Discriminate 3 test samples selected in Step 1 by the scoring system designed in Step 8.

Step 10. Repeat Steps 8 and 9 independently 10 times with 10 different sets of the training and test samples.

Fig. 2

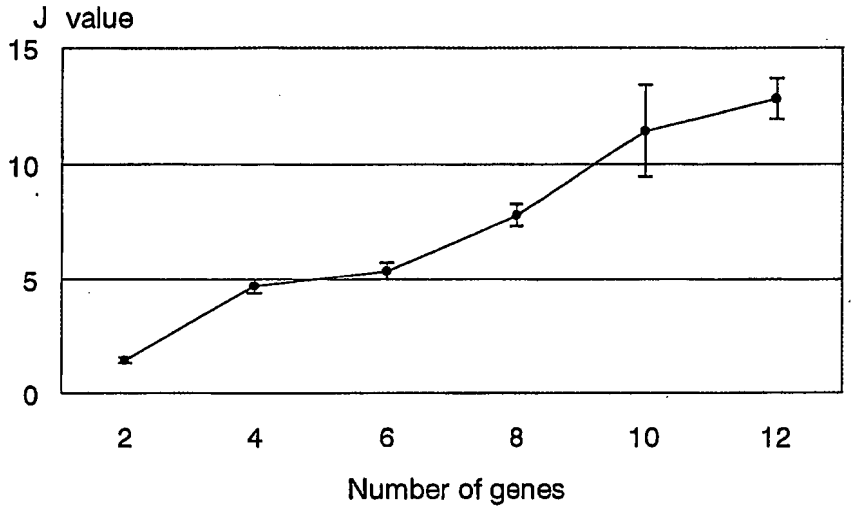


Fig. 3

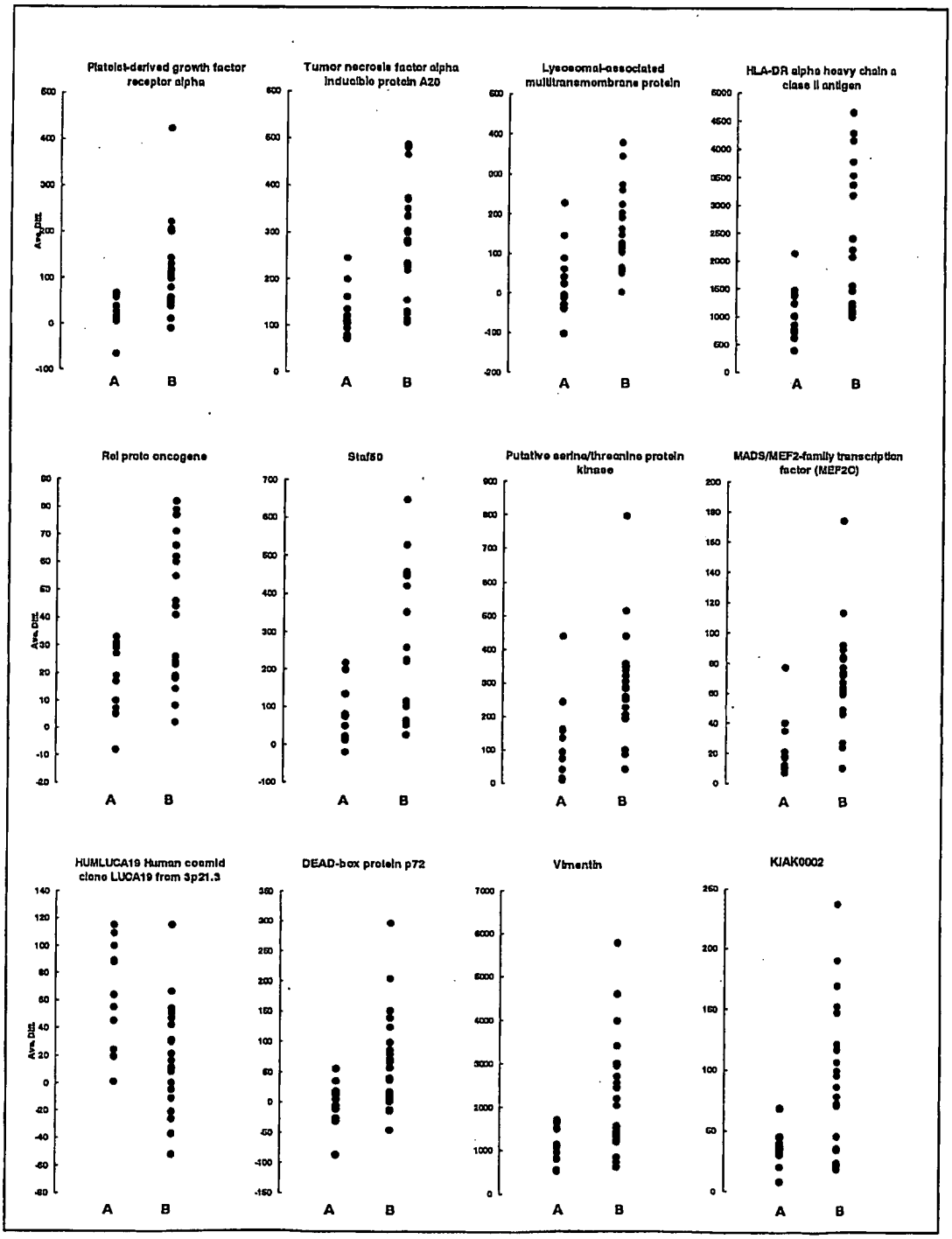

Platelet-derived growth factor receptor alpha
Tumor necrosis factor alpha inducible protein A20
Lysosomal-associated multitransmembrane protein
HLA-DR alpha heavy chain a class II antigen
Rel proto oncogene
Stat50
Putative serine/threonine protein kinase
MADS/MEF2-family transcription factor (MEF2C)
HUMLUCA19 Human cosmid clone LUCA19 from 3p21.3
DEAD-box protein p72
Vimentin
KIAK0002
Ave Diff
A
B

Fig. 4

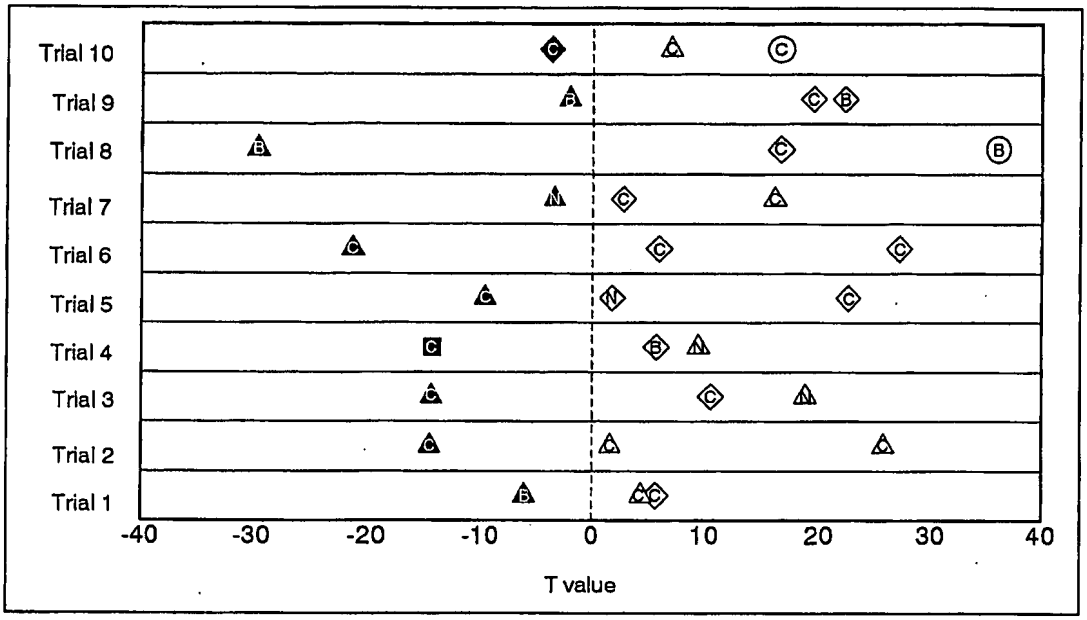
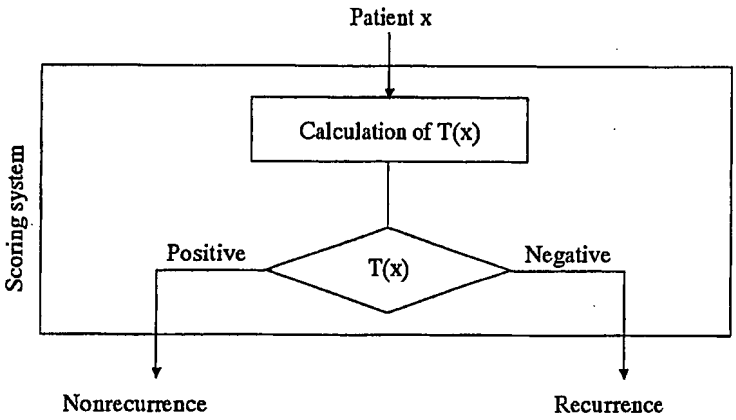

Trial 10
Trial 9
Trial 8
Trial 7
Trial 6
Trial 5
Trial 4
Trial 3
Trial 2
Trial 1
-40
-30
-20
-10
0
10
20
30
40
T value

Fig. 5


Patient x
Calculation of T(x)
Scoring system
Positive
T(x)
Negative
Nonrecurrence
Recurrence

INTERNATIONAL SEARCH REPORT

International Application No
PCT/JP 01/06330

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 C12Q1/68 G06F19/00 G01N33/574

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 C12Q G06F G01N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, BIOSIS, PAJ, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WO 98 56953 A (LEVINE MICHAEL A ;RINGEL MATTHEW D (US); UNIV JOHNS HOPKINS MED (U) 17 December 1998 (1998-12-17) claims 30-66	1-8
Y	US 5 862 304 A (MCGUIRE WILLIAM L ET AL) 19 January 1999 (1999-01-19) the whole document	1-8
Y	US 6 025 128 A (PARTIN ALAN W ET AL) 15 February 2000 (2000-02-15) claims 1,16-31	1-8
Y	DE 198 40 671 A (STIFTUNG TUMORBANK BASEL) 2 March 2000 (2000-03-02) claims 1-9	1-8

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
18 March 2002	22/03/2002

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Osborne, H

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WO 01 33228 A (FRUEHAUF JOHN ;ONCOTECH INC (US); MECHETNER EUGENE (US)) 10 May 2001 (2001-05-10) page 10, line 1 - line 8	1-8
Y	WO 01 18542 A (LEE JOHN ;LILLIE JAMES (US); THOMPSHO PAMELA (US); MILLENNIUM PRED) 15 March 2001 (2001-03-15) page 7, line 17 - line 26; claims 1-43,54-61	1-8
A	WO 97 09925 A (JACOBS IAN ;GEN HOSPITAL CORP (US)) 20 March 1997 (1997-03-20) claims 1-31	1-8
A	WO 01 40517 A (OXO CHEMIE AG ;MEUER STEFAN (DE); KUHNE FREDERICH WILHELM (TH); MC) 7 June 2001 (2001-06-07) the whole document	1-8
A	WO 01 00083 A (THOMAS JOEL ;INTERCET LTD (US); MEAGHER JOHN F (US); THOMAS AUSTIN) 4 January 2001 (2001-01-04) claims 1-5	1-8
A	JAIN A K ET AL: "STATISTICAL PATTERN RECOGNITION: A REVIEW" IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE INC. NEW YORK, US, vol. 22, no. 1, January 2000 (2000-01), pages 4-37, XP000936788 ISSN: 0162-8828 cited in the application the whole document	1-8

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/JP 01/06330

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9856953 A	17-12-1998	AU 7830598 A	30-12-1998
		WO 9856953 A1	17-12-1998
US 5862304 A	19-01-1999	AU 7980691 A	10-12-1991
		WO 9118364 A1	28-11-1991
US 6025128 A	15-02-2000	US 5989811 A	23-11-1999
DE 19840671 A	02-03-2000	DE 19840671 A1	02-03-2000
WO 0133228 A	10-05-2001	AU 1465801 A	14-05-2001
		WO 0133228 A2	10-05-2001
WO 0118542 A	15-03-2001	AU 7701100 A	10-04-2001
		WO 0118542 A2	15-03-2001
WO 9709925 A	20-03-1997	US 5800347 A	01-09-1998
		CA 2229427 A1	20-03-1997
		EP 0957749 A2	24-11-1999
		WO 9709925 A2	20-03-1997
		US 6030341 A	29-02-2000
WO 0140517 A	07-06-2001	AU 1806701 A	12-06-2001
		WO 0140517 A2	07-06-2001
		US 2001036631 A1	01-11-2001
WO 0100083 A	04-01-2001	AU 5897600 A	31-01-2001
		WO 0100083 A1	04-01-2001